RAKING RATIO ESTIMATION OVER TIME

Jeri M. Mulrow and H. Lock Oh
Internal Revenue Service

INTRODUCTION

Data from sample surveys are often analyzed using contingency table analysis. Raking ratio estimation is an iterative procedure which can be used to adjust contingency tables so that the marginal distributions obtained from outside sources are preserved. Raking is used at the Statistics of Income Division of the Internal Revenue Service in the SOI Corporate Study. In particular, this method is used during a post-stratification process to produce better industry estimates using population industry information. The iterative procedure that is used was first proposed by Deming and Stephan (1940) in connection with the 1940 U.S. Census.

There has continued to be interest in raking estimators throughout the years since 1940 from both a practical and theoretical point of view. Ireland and Kullback (1968) showed that raking estimates, although not maximum likelihood estimates under certain conditions, are consistent and best asymptotically normal (BAN). Brackstone and Rao (1979) developed asymptotic variance formulas for raking ratio estimators for up to four iterations. Causey (1983, 1984) continued to study raking estimates when data are assumed to be a random sample from a target population with known row and column margins. Bankier (1983, 1986) extended the work of Brackstone and Rao and developed a recursive method for computing asymptotic variances for an arbitrary, finite, number of iterations and for estimators of parameters that are defined by implicit functions. Little and Wu (1991) evaluated raking estimates when the sampled population differs systematically from the target population. They found that, in general, raking estimators compared favorably to the others reviewed in the paper.

Many Federal government agencies use raking techniques to produce estimates including, the U.S. Bureau of the Census, the Federal Reserve Board, and the Internal Revenue Service (IRS). The studies include the Survey of Consumer Finances, the Statistics of Income (SOI) Corporate Sample, and the U.S. census.

IRS has been using raking procedures since 1980 in the corporate study, incorporating a modification called bounded raking ratio estimation, see Leszcz, Oh and Scheuren (1983), Oh and Scheuren (1987), and Mulrow, Oh, and Collins (1991). This paper continues to explore the raking procedures developed in the SOI programs. In particular, it examines how stable the raking estimates have been over the last three years and the possibility of using raking to project ahead for future estimates.

We begin with a general description of raking in the SOI corporate program. A full statement of the problem addressed in this paper comes next. Then, stability of the raking estimates over time is explored next. This is followed by a section on the use of raking for projections. We end by presenting our initial results with some conclusions and a discussion of future work.

RAKING RATIO ESTIMATION

Raking ratio estimation usually assumes that two (or more) marginal population totals, say $N_{i.}$ and $N_{.j}$ are known, but the interior of the table N_{ij} can only be estimated from the sample by, say $\check{N}_{ij}$. In simple random sampling, the raking algorithm begins by setting

$$\check{N}_{ij} = (N/n) n_{ij},$$

and then proceeds by proportionately scaling the $\check{N}_{ij}$ such that the relations

$$\sum_J (\check{N}_{ij}) = N_{i.}$$

and

$$\sum_I (\check{N}_{ij}) = N_{.j}$$

are satisfied in turn. Each step in the algorithm begins with the results of the previous step, with the $\check{N}_{ij}$ continuing to change. The process terminates either after a fixed number of steps or when the sums are simultaneously satisfied to the closeness desired. (See Oh and Scheuren (1987) for further details.)

Typically, in survey sampling, we are interested in estimating the population total of a particular survey characteristic Y where

$$Y = \sum_{ijk} (Y_{ijk})$$

with the statistic

$$\hat{Y} = \sum_{ij} [(\check{N}_{ij} / n_{ij}) \sum_k Y_{ijk}].$$

In this setting, a raking weight

$$\hat{W}_{ij} = \check{N}_{ij} / n_{ij}$$

is placed on each individual record on the file for ease of handling. It is important to note that a feature of the raking algorithm is that if $n_{ij} = 0$ then necessarily $\check{N}_{ij} = 0$. For convenience, let $\hat{W}_{ij} = 0$ in such cases as well.

The situation in the SOI corporate sample is slightly different than that described above. In our setting the marginal population totals $N_{i.}$ and $N_{.j}$ are known and the interior of the table N_{ij} are also known. Raking is used as a way to systematically handle cells in the table where the n_{ij} are small or even zero.

In order to minimize the conditional bias, the conventional simple ratio estimator is used in "large" cells; thus a hybrid estimation method is actually used in the SOI corporate sample. The simple ratio estimator is used for cells with n_{ij} large. These cells are then removed from the population and sample tables, and the remaining sample cells are raked to the adjusted population marginals. Bounded raking is then used so that the weights, $\hat{W}_{ij}$, do not vary "too much" from the original weights. (See Mulrow, Oh, and Collins (1991) for further details.) Finally, an averaging procedure is used that guarantees that the sample cell size is greater than or equal to 25 for any given bounded raked weight.

STATEMENT OF THE PROBLEM

The above procedures have been used in the SOI corporate sample since 1984. Beginning with the SOI 1990 file, the main data users have requested advance data estimates which requires us to make projections of N_{ij}, the population cell sizes. The questions that thus arise are (1) How stable have these raked estimates been over the years? and (2) Can raking estimates be used over time to make projections of the N_{ij}? This paper will attempt to answer the first question and to present the initial results of procedures used to answer the second question. Before discussing the answers, it might be worthwhile to give some of the details of the SOI corporate sample. Those readers familiar with the sample may want to skip to the next section immediately.

Briefly, the SOI corporate sample is a highly stratified probability sample. It is first broken down by tax form type: 1120, 1120-A, 1120S, 1120RIC, 1120REIT, 1120F, 1120L and 1120PC. Within each form type, the returns are stratified by size. The definition of size varies with form type but is generally based on the income and/or asset size of the corporation. The sampling rates vary by

size and form type, but the larger corporations are always selected with higher probability than the smaller ones with the very largest corporations selected at 100%. Thus, the SOI corporate sample is skewed towards the larger corporations. (For a more detailed description of the corporate sample design and sampling rates see Mulrow (1992).)

Post-stratification by industry is performed for form types 1120, 1120-A and 1120S. There are fifty-eight original major industry groupings. It is often necessary to collapse industry groupings due to sparsity of the data. The collapsing scheme changes from year to year along with the definition of "large" for those cells which receive simple ratio estimates. Figure 1 shows how many industry grouping were collapsed and the definition of large for the three SOI years under consideration in this paper, 1987, 1988, and 1989.

Figure 1.--Post-stratification Details

SOI Year	"Large" n_{ij}	# collapsed industries	# raked cells
1987	300	13	460
1988	300	19	441
1989	350	20	433

Only three years of SOI corporate data are used to investigate the stability of the raking estimates over time due to the dynamic nature of tax law changes. Although data is available for many years back, a major revision of the tax laws was imposed in 1985. Data prior to this change may not be comparable to data after the change. Thus data prior to this change is not used in this investigation. In order to give taxpayers time to adjust to the changes, data from 1986 is not used in this investigation either.

It is necessary to investigate the stability of the raking estimates over time before a procedure can be developed to use raking to make projections. The next section discusses the stability question. A first attempt at using raking to make projections is presented in section five.

STABILITY

A minimum discrimination information number, $\mathbb{I}$, is used to measure the distance between the matrices of raked estimates for two years at a time. That is, to compare the stability of the raked estimates from 1988 to 1989, $\mathbb{I}$ can be written as follows:

$$\mathbb{I}(\check{N}^{89}, \check{N}^{88}) = 2 \sum_{ij} \check{N}_{ij}^{89} \ln(\check{N}_{ij}^{89} / \check{N}_{ij}^{88}).$$

Using properties of logarithms and usual raking techniques, the contribution to $\mathbb{I}$ due to yearly changes in the marginals and interior can be estimated separately. For example:

$$\ln(\check{N}_{ij}^{89} / \check{N}_{ij}^{88}) = \ln \check{N}_{ij}^{89} - \ln N_{ij}^{88}$$

$$= (\mu^{89} + \tau_i^{89,I} + \tau_j^{89,J} + \tau_{ij}^{89,IJ}) -$$

$$(\mu^{88} + \tau_i^{88,I} + \tau_j^{88,J} + \tau_{ij}^{88,IJ}),$$

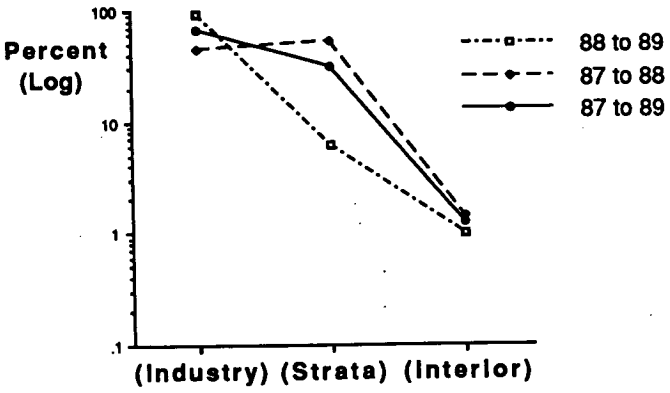
where

τ is the coefficient for factor i, j, or ij.

By fixing the various marginals for each year and employing raking, the contributions from each factor for each year can be estimated. Thus, changes in the raking estimates from year to year can be attributed to either changes in the marginal distributions of the sample strata sizes or the major industry grouping sizes or to changes in the interior of the table.

Figure 2 is a graphical representation of the percent of contribution to $\mathbb{I}$ due to these three factors. Comparisons are given for years 1988 to 1989, 1987 to 1988, and 1987 to 1989. Changes in the marginal distribution of major industry groupings was the major contributor to the value

Figure 2.--Percent Contribution to $\mathbb{I}$



of $\mathbb{I}$ in two out of the three comparisons. This does not come as a surprise, since the industrial groupings are part of the post-stratification process and the collapsing schemes changed from year to year. Differences in the interiors of the tables was minimal once changes for the marginal distributions were accounted for. That is, the raked estimates of the $\check{N}$ are fairly stable once changes in the marginal distributions of the sample strata and industry groupings are accounted for from year to year.

We are now confident that the stability of the estimates depends heavily on the marginal distributions and not on random variations in the interior of the table or the raking procedures. Therefore, if decent estimates of the marginals are known, raking may be able to be used to make projections. Figure 3 shows a comparison of the estimated population size strata counts to the actual population counts. Estimates of size strata counts were made four months in advance of final counts.

Although direct population estimates of the industry strata marginal counts are not available, we developed a method to use the prior year values with current year estimates. The next section describes in detail our first attempts at using raking to make projections.

Figure 3.--Population Estimates vs Actuals for Size Strata

STRATA	ESTIMATE	ACTUAL	% CHANGE
1	1627764	1636263	0.52
2	576906	579149	0.39
3	698253	701405	0.45
4	443433	444664	0.28
5	283624	284382	0.27
6	208428	208631	0.10
7	80356	80437	0.10
8	42420	42466	0.11
9	29735	27896	-0.14
10	12239	12317	0.63

RAKING FOR PROJECTIONS

Our estimates for the population size strata counts at advance data time were very good as can be seen from Figure 3. Thus our weighting procedures based on these values would seem to be fairly good. However, no industry adjustments or weighting techniques can be applied without further having good estimates of the industry strata marginal counts and interior counts. We concentrated our analysis on using raking to produce estimates of the N_{ij} and thus estimates of the $N_{i.}$ industry strata counts.

Full data is available for SOI years 1987, 1988 and 1989. Taking 1989 as the target year, then projections can be made using full information from 1987 and 1988 and partial information from 1989. A comparison of raked projections to the actual values for 1989 can then be made.

Assume the following partial data is available from 1989:

1) The sample counts by size and industry, n_{ij},
2) The marginal sample counts by size, $n_{.j}$,

3) The marginal sample counts by industry, $n_{i.}$; and
4) The estimated population counts by size, $\check{N}_{.j}$.

Then a simple estimate of the population counts by size and industry can be obtained from

$$\check{N}_{ij} = \hat{W}_{.j}\, n_{ij}$$

where

$$\hat{W}_{.j} = \check{N}_{.j} / n_{.j}.$$

Incorporating prior year information into the projection method could be accomplished by using the following "Stein-like" estimate:

$$\tilde{N}_{ij} = \alpha[\check{N}_{ij}^{PY}] + (1 - \alpha)[\check{N}_{ij}^{CY}]$$

where

$$0 \leq \alpha \leq 1,$$

$\check{N}_{ij}^{PY}$ is a raked prior year estimate, and

$\check{N}_{ij}^{CY}$ is the current year simple estimate.

Varying the level of α allows us to put more emphasis on either this year's simple estimate or last year's raked estimate. A value of $\alpha = .5$ puts equal emphasis on both estimates.

Thinking along these lines, lead us to further partition the data. If more emphasis could be placed on last years raked estimate than this years simple estimate, then in certain instances one may be better than another for a subset of the data. We looked at the following three cases to subdivide the information further:

Case 1: Marginal industry population counts $\geq$ 20,000 or not.

Case 2: Sample cell counts $\geq$ 50 or not.

Case 3: Sample cell counts $\geq$ 300 or not.

That is, large cells or large industries may be better estimated differently than smaller ones. The corresponding final estimate of $\tilde{N}_{ij}$ using Case 3 as an example is:

$$\tilde{N}_{ij} = \alpha[\check{N}_{ij}^{PY}] + (1 - \alpha)[\check{N}_{ij}^{CY}] \quad \text{if } n_{.j}^{CY} \geq 300$$

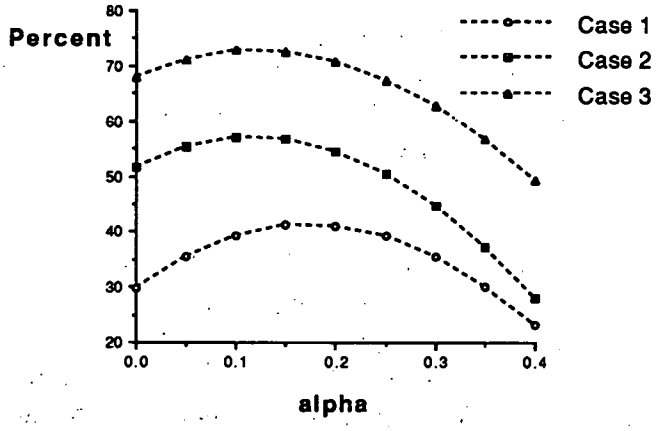
$$= \beta[\check{N}_{ij}^{PY}] + (1 - \beta)[\check{N}_{ij}^{CY}] \quad \text{if } n_{.j}^{CY} < 300$$

where

$$0 \leq \alpha \leq 1 \text{ and } \beta = 1 - \alpha.$$

In order to evaluate the estimates, we again looked at a minimum discrimination information number, $\hat{I}$, to measure the distance between the tables of estimates relative to the actual numbers. Figure 4 shows for different values of α (and thus β) how well subdividing the data and using the "Stein-like" estimate compared to just using the simple estimate from the current year. That is, how much closer to the actual numbers are the "Stein-like" estimates compared to the simple estimates. Figure 4 uses only prior year data from 1988 only. Using a two year lag in prior year information did not amount to any appreciable difference in the conclusions about the estimates.

Figure 4.--Percent Reduction in $\hat{I}$ Using 1987 and 1989 Data



As can be seen from Figure 4, subdividing the data based on the sample cell sizes produced the

greatest reduction in $\hat{I}$. Almost a 72% reduction in $\hat{I}$ can be achieved if $\alpha = 0.1$ and the data is subdivided based on sample cell size, n_{ij}. Thus, placing more emphasis on using the current year estimate when the sample cell sizes are large ($n_{ij} \geq 300$) and placing more emphasis on the prior year estimate when the sample cell sizes are small ($n_{ij} < 300$) results in an estimate of $\tilde{N}_{ij}$ that is 72% closer to the actual value than just using the simple estimate from the current year.

All three different subdivisions resulted in closer estimates than just using the simple estimate. The next steps are obvious. We need to consider more ways to subdivide. We need to vary the α levels more and take off the restriction that $\beta = 1$. We need to expand our "Stein-like" estimate to include more than one year of prior data. That is, can two or more years worth of prior data result in even better estimates or projections? A final model needs to be derived and checked against future data. Variance estimates need to be derived and a tie-in with bounded raking techniques pursued.

The initial work in this paper is far short of coming up with a final method for using raked estimates over time to make projections. The results, however, are promising. The ultimate goal is to be able to use such a method to help SOI produce data estimates on demand at any given timepoint in the sampling. Although this is an ambitious goal, it is nevertheless the one we are aiming for.

ACKNOWLEDGMENTS

The authors would like to thank Fritz Scheuren for his many ideas which lead to this work and to Beth Kilss and Wendy Alvey for helping in the preparation of this paper for both publication and presentation.

REFERENCES

BANKIER, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *JASA*, 81, 1074-1079.

BINDER, D.A. and THÉBERGE,A. (1988). Estimating the variance of raking-ratio estimators. *The Canadian Journal of Statistics*, Vol. 16, Supp. 47-55.

BRACKSTONE, G.J., and RAO J.N.K. (1979). An investigation of raking ratio estimators. Sankhyā, Ser. C, 41, 97-114.

CAUSEY, B.C. (1983). Estimation of proportions for multinomial contingency tables subject to known marginal constraints, *Communications in Statistics-Theory and Methods*, 12, 2581-2587.

CAUSEY, B.C. (1984). Estimation under generalized sampling of cell proportions for contingency tables subject to marginal constraints, *Communications in Statistics-Theory and Methods*, 13, 2487-2494.

DEMING, W.E., and STEPHEN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known, *Annals Mathematical Statistics*, 11, 427-444.

IRELAND, C.T., and KULLBACK, S. (1986). Contingency tables with given marginals, *Biometrika*, 55, 170-188.

KENNICKELL, A.B. and WOODBURN, R.L. (1992). Estimation of Household Net Worth Using Model-Based and Design-Based Weights: Evidence from the 1989 Survey of Consumer Finances, Working Paper, Board of Governors of the Federal Reserve System.

LESZCZ, M., OH, H. L., and SCHEUREN, F. (1983). Modified raking estimation in the corporate SOI program. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 107-111.

LITTLE, R. J. A., and WU, M. (1991). Models for contingency tables with known margins when target and sampled populations differ, *JASA*, 86, 413, 87-95.

MULROW, J.M. (1992). Sample Description and Data Limitations. U.S. Department of Treasury *Statistics of Income - 1989 Corporation Income Tax Returns*, Publication 16, Internal Revenue Service.

MULROW, J.M., OH, H.L., and COLLINS, R. (1991). An evaluation of bounded raking ratio estimation in the Statistics of Income corporate programs, *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 597-601.

OH, H.L. and SCHEUREN, F. (1987). Modified raking ratio estimation. *Survey Methodology*, 13, 2, 209-219.